SKA大数据的科学应用和挑战

安 涛[1*] 武向平[1,2] 洪晓瑜[1] 叶叔华[1] 毛羽丰[3] 郭绍光[1] 劳保强[1]

1 中国科学院上海天文台 上海 200030

2 中国科学院国家天文台 北京 100101

3 中国科学院 前沿科学与教育局 北京 100864

摘要 即将开建的平方公里阵列（SKA）射电望远镜是最大的天文观测装置，有望在宇宙起源、生命起源、宇宙磁场起源、引力本质、地外文明等自然科学重大前沿问题上取得革命性的突破。SKA的超级灵敏度、超大视场、超快巡天速度和超高时间、空间、频率分辨率等技术特点确保了SKA在观测能力上的领先地位，由此也产生了海量观测数据。SKA的数据运输、存储、读写、运算、管理、归档、发布对信息和计算机领域的前沿技术均提出了严峻的考验。中国SKA科学团队将协同信息产业界一道应对SKA大数据的挑战，不仅推动产生重大原创性科学发现，其技术成果也将应用于国民经济建设。

关键词 平方公里阵列，大数据，高性能计算，科学应用

DOI 10.16418/j.issn.1000-3045.2018.08.016

天文学是一门最古老的学科，伴随着人类文明产生，而中国则是世界上天文学起步最早的国家之一。现代观测天文学从伽利略发明天文望远镜算起，至今已经有 400 多年的历史，天文学的每一次重大进展都离不开天文望远镜能力的飞跃式进步。

中国正处在新时代科技创新的战略机遇期，国家对科研的投入达到前所未有的高度。仰望星空离不开精密望远镜，近几年一批大型天文望远镜在我国相继建成，如兴隆大天区面积多目标光纤光谱天文望远镜（LAMOST）、贵州 500 米口径球面射电望远镜（FAST）、暗物质探测空间望远镜“悟空”、硬 X 射线调制望远镜“慧眼”，这些设备接近或达到国际一流水平。中国参加了世界上最大的天文大科学工程——平方公里阵列（Square Kilometre Array，SKA）射电望远镜的国际合作，其建成后将成为射电望远镜中的旗舰，树立自然科学探索历程中新的里程碑。现有的望远镜设备也都在升级更新，形成了从地基设备到空间卫星（以及空间站）的观测条件和从 X 射线、紫外线、光学、红外到

*通讯作者

资助项目：科技部国家重点研发计划大科学装置专项（2018YFA0404600），中国科学院国际合作局国际大科学计划培育专项（114231KYSB20170003）

修改稿收到日期：2018年8月11日

射电的全波段观测能力，把天文学的研究推上了指数增长的大数据时代。目前的天文数据已经达到了PB量级，随着观测技术的进步和观测设备的更新，很快将会进入到EB量级时代，天文大数据将深刻改变人类探索和认识自然的方式。

1 天文学研究已经步入大数据时代

从20世纪60年代以来，天文学不断产生令人赞叹的成果，天文学正书写着人类自然科学发展的辉煌篇章。最精彩、最具突破性的天文发现越来越依赖于大型科研装置的协同运行，越来越依赖于海量数据的分析和挖掘；同时，科学成果的透明度、多样性、多学科之间的融会贯通使得人类的科技生活越来越丰富多彩。天文学真正进入了多波段、多信使时代，人们不仅能够使用多个观测设备同时探测同一天体，获得几乎整个电磁波谱的完整信息，而且还能够使用电磁辐射之外的其他信源，比如中微子和引力波来研究宇宙天体。一个最具代表性的例子是2017年8月天文学家首次发现两颗中子星的并合事例。地基激光引力波天文台（LIGO）和VIRGO引力波探测器首先发现了中子星并合过程产生的时空涟漪，随后最强大的太空望远镜和地面望远镜协同观测并合后的后随辐射，使得人们不仅增进了对引力波的认识，而且从观测上证实了短伽马暴、巨超新星等奇异天体，这让我们对天文学协同研究的强大威力有了新的理解。

以观测为基础的天文学曾长期受到数据匮乏的困扰，进入21世纪信息时代，天文学已经发生了重大的革命性变化。天文观测已经逐步进入大数据时代，当前科学研究方式和传播方式也发生着深刻演变。举个例子：超新星是宇宙中绚烂的烟花，我国有世界公认的关于超新星的最早天文记录。超新星在天体物理研究中有重要的地位，2011年的诺贝尔物理学奖授予3位天文学家，他们的贡献是通过对超新星的观测发现宇宙正在加速膨胀。超新星是非常稀有的事件，在10年前捕获一颗超新星是相当困难的，因此每次观测到一个超新星也必然引起全球望远镜的追逐竞赛，大量研究不得不依赖于数值模拟和理论计算。而如今，光学巡天每年都能发现1 000多颗，超新星变得不再稀奇，深度有效地挖掘这些大型巡天积累的数据则有可能会产生更多新发现。随着SKA等下一代超级望远镜带来的天文观测能力的极大提升，在当前仍属于凤毛麟角的奇异天体在5—10年后都将成为常客。统计学、信息科学与天文学密切结合，为天文学家提供数据分析工具，基于对宇宙大数据的收集、整理、分析探索宏观宇宙的真理和天体的运行规律。

2 大数据典型应用——平方公里阵列（SKA）射电望远镜

天文学关注有关宇宙、天体和生命起源的最具有前瞻性的问题，这些问题的突破和解决将极大地推动自然科学基础理论，促进人类科技水平的整体进步。

由宏伟科学目标驱动的SKA射电望远镜是我国参加的最大的天文领域国际合作大科学工程。SKA建成后将成为世界上最大的天文实验装置，为人类探索宇宙起源奥秘创造新的机会。SKA由包括中国在内的11个正式成员国以及10多个观察员国参与，建设和运行天文大望远镜已经成为一个国家综合实力的真实体现和重要标志。SKA总部位于英国，SKA低频阵列（SKA-low）包括130万个对数周期天线，拟建于澳大利亚西部沙漠；SKA中频阵列（SKA-mid）包括2 500个碟形天线，拟建于南非以及南部非洲的无线电宁静区域，这两处是经过天文学家十几年评估和测评后优选出来的最佳台址。望远镜的总接收面积高达1平方公里，频率几乎连续覆盖50 MHz—20 GHz的范围，比目前厘米波段最大的射电望远镜阵的灵敏度提高约50倍、巡天速度提高约10 000倍[2]。

作为下一代担当引领作用的射电天文观测设施，SKA将对射电天文学的发展产生深远影响。SKA的强大

观测能力体现在其超高灵敏度（mK）、超大视场（数十度）、超快巡天速度、超高频率分辨率（kHz）、超高时间分辨率（纳秒）、超高空间分辨率（亚角秒），这些技术特点使得SKA产生前所未有的超大数据量[2]。

SKA的建设主要分为两个阶段：第一阶段（SKA1）将按照全规模的10%来建造，预计2020年开建；第二阶段（SKA2）将完成其余90%建设工程，不过目前尚未确定具体计划。SKA1-low每个台站的数据产生率为2 Tbps，总的数据流是1 Pb/s。据此规模递推，SKA2至少产生10倍以上的实时数据流。从上述数据可知，SKA产生的数据量是空前巨大的，即使经过相关处理后数据量极大降低了，但输入到科学数据处理器（SDP）①的数据也达到了4 GB/s，是当之无愧的科学大数据。SKA超大规模的数据流需要及时地以实时模式处理掉，否则会造成整个数据处理管线（pipeline）的堵塞甚至崩溃。采用实时模式、多并发任务、数据流管线系统的处理方式是SKA数据处理的几个典型特点，也是新型科学大数据处理的典型应用[3]。

作为史上最大的射电望远镜，SKA不仅承载孕育世界级科研成果的使命，而且将产生世界上最大规模的数据，因此我们需要充分认识到SKA数据处理的巨大挑战。由于SKA工程极其庞大及复杂，为了攻克关键技术、降低技术风险，包括中国在内的多个国家先后建设了一些探路者和先导项目，每个项目相当于SKA总体规模的1%左右，并基于这些先导望远镜开展了相关的科学预研究和技术攻关。这些先导设备在理解SKA科学目标、建立和逐步完善天空模型、开发和测试数据处理软件、培养急需的人才队伍等方面发挥了积极作用，在SKA发展历程中处于不可忽视的地位。需要指出的是，尽管如此，这些先导项目的数据量远远不能达到SKA1规模，因此与建立真实的验证参考还有一定的距离②。

3 SKA科学计算的挑战

与传统望远镜相比，SKA更像是一个“软件”望远镜，它不仅集成了当代信息计算技术的最新成就，而且提出了新的问题。以SKA-low为例，其旨在探测微弱宇宙信号，这些低频阵列以10 Pb/s速度产生出世界上最大规模的数据流，远远超出了全世界互联网的流量。按照SKA的数据流规模，估计在建设的第一阶段每年需要输送到区域数据中心进行深度分析的科学数据就达到了每年300 PB，随着望远镜的全面运行，这个数据量必然会逐步增加。到了SKA2阶段，从SKA天文台产生的预处理数据的规模将扩展到SKA先导项目的100倍以上，达到EB量级。SKA两个最重要的科学方向——宇宙再电离和黑暗时期探测、用脉冲星计时阵精确测量引力，需要积累未校准的原始数据；如果考虑到保存一定时间的原始数据，那么SKA天文台的数据存储需求将提高至少一个量级。

以SKA先导项目MWA为例，经过4年的运行，MWA积累了24 PB的科学存档数据。其中一个科学目标是GLEAM巡天任务，第一期巡天已经包含了30多万颗星系，存档数据量达到1 PB以上。第二期巡天已经开始，灵敏度提高了4倍以上，数据量预期高达6.5 PB。而MWA只占到SKA-low规模的1%，SKA数据量由此可见一斑。据初步估计，SKA1阶段的科学数据处理器所需要的计算能力为260 PFlops（即每秒260千万亿次浮点运算），相当于我国超级计算机“天河二号”的8倍，“神威·太湖之光”的3倍。SKA巨大的计算需求必然对现有科学计算的架构和方式形成巨大冲击，对SKA数据处理问题的解决有助于带动和提升相关产业的发展，甚至引发革命性变化。

SKA将对除天文学以外的其他众多学科诸如计算机科学、信息学、电子学等领域带来极大的促进作用③。

① 即建于两个台址国专门对这些原始科学数据进行预处理的超级计算机。
② 武向平，等。中国SKA科学白皮书（2017年）。
③ 武向平，等。中国参加SKA（第一阶段）综合论证报告（初稿），2018年。

TB量级的高速数字化采样、高速实时数字信号处理对电子行业带来新的挑战。长期工作在野外恶劣环境下射频信号长距离光纤传输的频率同步是孔径阵列急需解决的技术挑战之一。大数据长距离的高速宽带洲际传输对目前的科研网络基础设施、拓扑结构、通信协议、传输端软件等提出了严苛的要求——满足超高速流式数据处理设计的互联网络不是简单通过增加节点的互联网口数量和增加节点间的互联总带宽能实现的，对这个问题的有效解决也必将促进国内百GB甚至TB级基础网络的布局和建设。

以数据密集型科学计算为特点的SKA数据处理对我国的电子、计算机、信号处理行业提出了更高的要求。SKA科学数据处理应用面临着“存储墙”问题，即I/O问题，传输带宽是主要的系统瓶颈之一。即使“天河二号”这样的超算对于SKA这类大数据的处理资源也会有不足，同时不便进行突发事件的观测分析，因此亟待研究适应数据密集型科学计算的新型架构体系[4]。前面讲到，SKA高速海量的输入数据必须通过实时处理降低后续流程的压力，海量数据实时处理对软硬件体系都有特殊设计要求，整个系统的架构设计和集成安装、超算中心的执行框架和配套软件算法、数据中心的健康监控、机柜冷却、总控管理等都会面临诸多挑战；而且在建设经费封顶的情况下，既要达到预定的运算能力和实时性要求，还要从运行成本上考虑满足低功耗的要求。此外海量数据的存储、归档、检索、运算对超级计算机的完整生态链提出了极高的要求。尽管国产CPU芯片已经部署在国内大型超算中心，国内科研单位也研发了用于人工智能领域的深度学习处理器芯片；但不容乐观的是，目前主流的操作系统、存储系统等软件生态基本全部来自于国外，最关键的软件生态环境依然远远落后国际水平，尚不具备竞争力，“卡脖子”问题依然严重，自给自足的能力还不够。SKA项目为相关产业的发展提出了强烈的需求驱动。

除了硬件方面的问题，天文应用软件的目前研发水平也远远无法达到SKA的要求。SKA科学数据处理的关键算法存在大量对共享资源包括共享文件系统的操作，传统固定多核的计算机系统在多任务、多并发、多线程并行执行时经常出现资源竞争；如果数据流执行框架不能有效地妥善解决资源调度和分配，严重的情况下将导致数据处理流水线停顿[3,4]。实际上，这一问题在SKA先导望远镜数据处理中心并不罕见。为此，澳大利亚ICRAR研究所和中国科学院上海天文台针对SKA项目联合研发了名为*Data Activated* 流（*Liu*）*Graph Engine*（DALiuGE）的数据流执行框架[3]，其采用了“数据驱动”的先进设计理念，比传统的HPC“计算驱动”的设计更适合SKA科学计算。此外，SKA科学计算的实际运算效率小于原计划的10%，因此其原定理论峰值性能260 PFlops无法完成科学数据处理的实际需求。增加超算资源的简单做法并不切实可行，更加可行的途径是提高软件执行效率——效率从10%提高到20%，可以节约50%的计算资源以及大幅度降低运行成本。天文学家与计算机专家合作优化代码，可以数倍地提高算法和程序的运行速度。当务之急是培养既懂天文又懂计算的复合型人才。另一个现实的问题，天文数据处理的软件也亟待更新换代以满足未来的需求。目前主要使用的天文软件大部分在20世纪70—80年代研发，考虑到天文应用对高速、实时、并行的大数据处理需求，天文学家已经开始使用更先进、更模块化、支持并行的开发语言，如C++或者Python。使用C++开发的AIPS软件的替代版本CASA软件将成为下一代主流射电天文软件；涉及机器学习、人工智能的程序将以Python为优先选型。天文数据处理软件的研发与天文研究一样，已经从单打独斗模式升级为全球化合作集体作战，比如发现引力波的LIGO团队由1 000多位科学家组成，广泛应用于射电天文处理软件的CASA核心库也有来自全球近百位人员贡献代码及算法；航空母舰式的联合研究团队，大兵团作战模式将成为解决重大科学问题的标准资质。

科学传播比任何时候都得到重视，“科技创新、科

学普及是实现创新发展的两翼”[5]。未来 SKA 的天文大数据将不仅仅服务于天文学家，也将提供面向公众的接口。以 SKA 为依托，宣传科研成果、交流学术思想、普及科技知识、弘扬科学精神，大力推广基础科学在公众间的认知度，提高科研在公众的普及度。SKA 区域中心将通过虚拟天文台和“云”的方式让老百姓以更加便捷的方式接触科学，在公众中普及天文学。

4 中国 SKA 科学和区域中心的思考和对策

我国正面临着推进科技创新的重要历史机遇。科技创新已经被提升到实现“两个一百年”奋斗目标、实现中华民族伟大复兴的中国梦的战略高度。SKA 是我国参加的最大的天文领域国际合作项目，为我国射电天文学实现从“跟跑”到“并跑、领跑”创造了难得的机遇。SKA 将主导和影响未来 50 年射电天文学的发展命运，使低频射电天文学再次进入蓬勃发展的新时代，将孕育诸多重大科学突破，创造观测宇宙学研究的又一个辉煌④。

SKA 数据的深度分析和加工是在分布于几大洲的区域数据中心完成。包括中国在内的几个主要成员国对于建设 SKA 区域数据中心均予以积极态度并寄予很高的期待，已经开始了关键技术研究工作。由于 SKA 数据处理的特殊性、复杂性、巨大数据量，大规模的数据搬运是不现实的，因此中心化的数据处理方式成为必然选择。建设中国 SKA 区域中心不仅是国际 SKA 总体规划不可或缺的一个部分，也是支撑中国科学家有效利用 SKA 数据获得相应科学回报的重要保障。SKA 科学家在全球广泛分布，分布式计算和存储、云化成为数据存档和发布的考虑，多个科学和数据分中心组成的区域中心网格可以满足 SKA 的多样化需求。中国科学院上海天文台与澳大利亚的 SKA 数据中心之间已经建立了端对端的直连，最高数据传输速率达到 3.2 Gbps，是目前已知最高的天文数据流速率，这为 SKA 区域中心提供了有益经验和实际模版。SKA 多科学目标多种数据属性的特征使得多数据流并行成为必然趋势，也是未来 SKA 区域中心国际网络建设方面需要关注的问题。

为了与国际同步乃至赶超，依托 SKA 这样的大科学工程要顶层设计，定向规划人才培养，不仅要坚持独立自主，还要多与国际顶级研究单位、一流团队进行合作开展前沿研究，提升自身能力。中国目前严重缺乏数据处理人才，要充分认识到人才培养的长期性。中国 SKA 科学团队要抓住 SKA1 第一批数据发布（2022 年）之前这段宝贵的时间窗口，围绕相关的科学研究，利用 SKA 先导望远镜产出科学成果、掌握数据处理技术，争取在 SKA1 运行后能够尽快投入相关科学研究。除了天文研究和数据处理人才，在大型国际合作科技项目中，管理型科技专家要走上国际舞台，不断巩固和加强学术地位。科学家要勇于承担使命，争取在国际组织中担任重要职务，在国际大科学工程中掌握话语权，维护国家利益，配合民族复兴的国家战略。

应对 SKA 大数据的挑战，应一方面立足国际合作，另一方面加快关键核心技术国产化。可以考虑以中国 SKA 区域科学和数据中心为依托，争取突破 TB 量级高速科研骨干网、信号与数据传输以及 EB 量级高性能计算机等关键技术，开发出配套的天文软件来支持相应天文课题的数据处理，从而在 SKA 时代到来之际能够使用 SKA 科学数据快速取得重大科学成果，引领先进科学方向。

总之，人类共享一个天空，通过参与 SKA 全球创新合作，共同促进天文学的跨越式发展，为解决人类共同关注的科学目标做出贡献，是“构建人类命运共同体”理念的重要实践。

参考文献

1 SKA Organisation. Advancing Astrophysics with the Square

④ 叶叔华，武向平。我国低频射电天文学学科发展战略咨询研究报告。中国科学院学部咨询报告，2018 年。

Kilometre Array. Italy: Proceedings of Science (AASKA14), 2015: 1-1949.
2 Dewdney P E, Hall P J, Schilizzi R T, et al. The Square Kilometre Array. Proceedings of the IEEE, 2009, 97(8): 1482-1496.
3 Wu C, Tobar R, Vinsen K, et al. DALiuGE: A graph execution framework for harnessing the astronomical data deluge. Astronomy and Computing, 2017, 20: 1-15.
4 安涛. 超级望远镜的数据革命. 科技纵览, 2017, (11): 64-65.
5 习近平. 为建设世界科技强国而奋斗——在全国科技创新大会、两院院士大会、中国科协第九次全国代表大会上的讲话. [2016-05-30]. http://news.xinhuanet.com/politics/2016-05/31/c_1118965169.htm.

Science Applications and Challenges of SKA Big Data

AN Tao[1*] WU Xiangping[1,2] HONG Xiaoyu[1] YE Shuhua[1] MAO Yufeng[3] GUO Shaoguang[1] LAO Baoqiang[1]
（1 Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 200030, China;
2 National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China;
3 Bureau of Frontier Sciences and Education, Chinese Academy of Sciences, Beijing 100864, China）

Abstract The Square Kilometre Array (SKA) radio telescope to be built soon is the largest astronomical observing facility, and it is expected to make revolutionary breakthroughs in the major frontiers of natural sciences to answer fundamental questions of origins, such as the origin of the Universe, the origin of life, the origin of the cosmic magnetic field, the nature of gravity, and search for extraterrestrial civilization. The unprecedented power of the SKA, characterized by the extremely high sensitivity, wide field of view, ultra-fast survey speed, super high time, space, and frequency resolutions ensures the leading position of the SKA in radio astronomy in next decades, which also produces a vast amount of observational data at ExaByte (EB) level. The transportation, storage, reading, writing, computing, management, archiving of the SKA-level data and the release of SKA science products have posed serious challenges on the technologies in the field of information and computers. China SKA science team will work together with the information, communication, and computer industry to tackle the challenges of the SKA big data, as not only promotes major original scientific discoveries, but also applies the derived technological achievements to stimulate the national economy.

Keywords Square Kilometre Arra (SKA), big data, high performance computing, science application



安 涛 中国科学院上海天文台研究员。国际天文学会射电专业组委会委员，中国天文学会青年工作委员会副主任，上海天文台SKA团队课题组长。主要研究领域包括：射电天文，天体物理，天文技术与方法。E-mail: antao@shao.ac.cn

AN Tao Research Professor of Shanghai Astronomical Observatory (SHAO), Chinese Academy of Sciences (CAS). Organization Committee member of Commission B4 Radio Astronomy, International Astronomical Union, Deputy Director of Youth Committee, Chinese Astronomical Union, Head of SKA group, SHAO. Research interests include radio astronomy, astrophysics, and astronomical technique. E-mail: antao@shao.ac.cn

■责任编辑：岳凌生

*Corresponding author